农业病虫害图像数据集构建关键问题及评价方法综述

管博伦,张立平,朱静波,李闰枚,孔娟娟,汪 焱,董 伟*

(安徽省农业科学院农业经济与信息研究所,安徽合肥 230001,中国)

摘 要:［**目的/意义**］农业病虫害科学数据集是农业病虫害监测预警的基础,也是发展智慧农业重要的组成部分,对农业病虫害防治具有重要意义。随着深度学习技术在农业病虫害智能监测预警中应用效果的凸显,构建高质量的农业病虫害数据集逐步受到专家学者的重视。为了进一步构建高质量、分布均衡的农业病虫害图像数据集,提高检测模型的准确性和鲁棒性,本文以构建农业病虫害图像数据集面临的挑战为切入点,对农业病虫害数据集的构建进行了全面综述。［**进展**］分别从数据集层次、数据样本层次和使用层次总结构建农业病虫害图像数据集所面临的类间类内样本不均衡、选择偏差、目标多尺度、目标密集、数据分布不均、图像质量参差不齐、数据集规模不足以及数据集可用性等问题,从图像采集和标注方法两个方面,分析以上问题的主要成因,并归纳算法的改进策略和建议,最后总结了数据集相关评价方法。［**结论/展望**］结合农业病虫害图像识别实际需求,对构建高质量农业病虫害图像数据集提出了相关建议:(1)结合实际使用场景构建农业病虫害数据集。多视角、多环境下采集图像数据构建数据集,从算法提取特征的角度,科学、合理划分数据类别,构建样本数量分布和特征分布均衡的数据集;(2)平衡数据集与算法间的关系。研究数据集特征与算法性能之间的关系,需充分考虑数据集中的类别和分布,以及与模型匹配的数据集规模,以提高算法准确性、鲁棒性和实用性。深入研究农业病虫害图像数据规模与模型性能的关联关系、病虫害图像数据标注方法、模糊、密集、遮挡等目标的识别算法和高质量农业病虫害数据集评价指标,进一步提高农业病虫害智能化水平;(3)增强数据集的使用价值。构建多模态农业病虫害数据集,创新数据采集组织形式,开发数据中台,挖掘多模态数据间的关联性,提高数据使用便捷性,为应用落地、业务创新提供高效服务。

关键词:农业病虫害;数据集;深度学习;监测预警;数据采集;数据标注;数据集评价

中图分类号:S-1;TP392 **文献标志码:**A **文章编号:**SA202306012

引用格式:管博伦,张立平,朱静波,李闰枚,孔娟娟,汪焱,董伟.农业病虫害图像数据集构建关键问题及评价方法综述[J].智慧农业(中英文),2023,5(3):17-34. DOI:10.12133/j.smartag.SA202306012

GUAN Bolun, ZHANG Liping, ZHU Jingbo, LI Runmei, KONG Juanjuan, WANG Yan, DONG Wei. The key issues and evaluation methods for constructing agricultural pest and disease image datasets: A review[J]. Smart Agriculture, 2023, 5(3): 17-34. DOI:10.12133/j.smartag.SA202306012 (in Chinese with English abstract)

1 引 言

农业病虫害一直以来是影响农作物产量与质量的重要因素之一。据《农作物有害生物名录》记载,截至2014年,中国粮、棉、麻、油、糖、果、茶等发生的有害生物种类及其检验出有害生物种类3600多种,其中害虫2000余种、病害800余种、杂草680余种、鼠害66种[1];截至2013年,国内已知取食蔬菜的害虫2460种[2]。利用测报手段对农作物虫害进行有效监测可以减少农药的使用,保障农作物的质量和产量,起到降本增效的作用。农业病虫害的测报长期以来主要是以人工方式为主进行

收稿日期:2023-06-13

基金项目:国家自然科学基金面上项目(32171888);安徽省财政农业科技成果转化项目(2022ZH001);安徽省农业科学院科研计划项目(2023YL014)

作者简介:管博伦,研究方向为数据挖掘、机器学习、农业信息化。E-mail:aaasguanbolun@163.com

*通信作者:董 伟,博士,副研究员,研究方向为农业信息化。E-mail:dw06@163.com

现场识别与统计[3]，但因基层测报人员不足、专业水平参差不齐等因素，导致人工测报效率低、可靠性较差。近年来，随着农业物联网技术和计算机视觉技术的发展，利用信息化手段对农业病虫害进行智能识别与精准防治成为可能，也是智能农业在病虫害精准识别应用领域的必然发展趋势，它的应用场景非常广阔，包括但不限于农作物病虫害识别、农业病虫害监测预警、农作物目标检测、农作物受害程度评估、农作物产量预测等[4-6]，具有节省时间、减少人为主观性和增加安全性等优点[7]。信息化技术需要具有多尺度信息且构建科学的高质量数据集作为支撑[8]，合理科学地构建农业病虫害数据集至关重要，它是农业病虫害智能识别的重要组成部分，也是病虫害监测预警的基础。

影响图像识别效果的因素之一是数据集质量。通常在进行图像识别研究时，往往专注于模型本身。事实上，随着大数据时代的到来，在图像识别任务中数据的作用越来越明显，很多研究者也意识到了数据的重要性，开始关注数据质量的高低[9]，2019年由科技部认定的国家农业科学数据中心为整合农业科学数据资源、共享农业数据资源发挥了重要作用。农业病虫害智能识别是以病虫害数据集为核心展开的，许多专家在农作物虫害图像数据集构建方面进行了研究。De Cesaro Júnior和Rieder[10]分析了害虫图片中难以解决的遮挡问题；Li等[11]通过介绍害虫分类的应用，总结了图像采集和预处理等方法，并提出一个农业害虫检测系统结构；汪京京等[12]介绍了农作物虫害图片中分割和特征提取的方法，阐述了农作物病虫害识别算法的进展；翟肇裕等[3]从虫害数据的获取、数据的处理以及数据的应用三个方面介绍了虫害识别的关键技术；Hasan等[13]指出，在虫害识别过程中，样本图像具有目标小、目标密集等特征，这些特征已经成为虫害识别的主要问题之一，并针对性地提出相关算法。上述文献均针对虫害识别从数据集到识别算法上进行探讨，侧重于识别算法的研究。本文针对虫害识别中数据集的构建，聚焦于农业病虫害识别方法中科学构建数据集的角度进行综述；分析了构建病虫害数据集的挑战，并从采集和标注过程两个方面分析问题产生的原因以及相关解决方法；进一步总结了病虫害数据集的评价方法。

2 农业病虫害数据集建设现状

2.1 常见农业病虫害数据集

病虫害图像识别技术是深度学习中的图像识别算法在农业方面的应用，深度学习算法依赖于科学合理的数据集，算法对构建的数据集质量有一定的要求[14]。一个标注准确、规模大小适当、种类样本均衡、高相关性的数据集对模型算法的训练和测试的准确性，以及实际使用的效果好坏能够起到举足轻重的作用[15]。农业病虫害数据集主要来源方式有两种：一种是根据实际需求自己构建的私有数据集，其特征一般是包含病虫害种类较少、类间样本较多、图像质量高、标注正确性较高，但是不公开；另一种是网络上开源的公共数据集，其特征一般包含的病虫害种类较多、类间样本重复率高、图像质量较低、标注正确性较低，但是可以公开使用。表1为一些农业病虫害相关的数据集。

表1 不同农业病虫害数据集对比

Table 1 Comparison of different agricultural pest and disease datasets

序号	数据集名称	类别数量/个	描述	来源
1	Plant Leaves[16]	22	覆盖12种植物，包括芒果，阿琼，雪桐，番石榴，白耳，贾蒙，麻风树，蓬蓬，罗勒，石榴，柠檬和中国芹等植物，共4503张图像，2278张健康的叶片和2225张患病的叶片	https://www.kaggle.com/datasets/csafrit2/plant-leaves-for-image-classification
2	Plant Village[17]	38	利用互联网图像对健康和病害的作物叶片进行标注，一共38个类别，涵盖了苹果、蓝莓、玉米、葡萄、橘子等作物以及作物的17种真菌疾病、4种细菌疾病、2种霉菌疾病、2种病毒性疾病、1种由螨引起的疾病共54,303张健康和病害图片	https://github. com/spMohanty/PlantVillage-Dataset

续表

序号	数据集名称	类别数量/个	描述	来源
3	IP102[18]	102	主要在互联网上搜集图片并进行标注形成的数据集，含有幼虫、成虫等不同的形态的102个害虫类别。共75,222张图像，训练集45,095张，验证集7508张，测试集22,619张	https://github.com/xpwu95/IP102
4	Rice Leaf Disease Images[19]	4	作者自行拍摄构建的患病水稻叶片图像数据集，使用尼康DSLR-D5600拍摄，部分样本来自网络图像，单张图像像素大小为300×300。包含细菌性枯病、稻瘟病、褐斑和苔斑4种，共5932张图片，其中测试集800张，5132张被增强用作训练集	https://doi.org/10.1016/j.compag.2020.105527
5	大田作物病害识别研究图像数据集[20]	15	以图像数据库的形式存储，包含小麦、水稻、玉米3种大田作物的15种病害，共17,625张样本	http://www.doi.org/10.11922/sciencedb.745
6	葡萄病害识别图像数据集[21]	7	包含葡萄白粉病、葡萄花叶病毒病、葡萄黑霉病、葡萄灰霉病、葡萄溃疡病、葡萄霜霉病和葡萄酸腐病7种病害，共3622张样本	http://www.doi.org/10.11922/sciencedb.j00001.00311
7	AgriPest[22]	14	共49,707张图像样本，大概按照9:1的方式划分为44,716张训练数据集和4991张验证数据集，包含4种作物的14类害虫	https://www.mdpi.com/1424-8220/21/5/1601
8	苹果叶片病害数据集[23]	5	包含斑点叶落病、褐斑病、花叶病、灰斑病和锈病5种病害，原始图片2029张，其中411张落叶病、435张褐斑病、375张花叶病、370张灰斑病和438张锈病。数据增强后共24,348张样本，图像像素大小统一为512×512	http://www.agridata.cn/data.html#/paperdetail? id=4363
9	桉树害虫数据集[24]	3	桉树红胶木虱（Eucalyptus redgum lerp psyllid，Glycaspis brimblecombei）、桉树桐蝽科害虫（haumastocoris peregrinus）和一种寄生虫，共748张样本，图像像素为500×500	https://www.frontiersin.org/articles/10.3389/fevo.2021.600931/full
10	Rustia2021[25]	4	包含苍蝇、蓟马、粉虱、蠓类4种虫害，共990张样本	https://onlinelibrary.wiley.com/doi/10.1111/jen.12834
11	Pest24[26]	24	包含24种害虫，共25,378张样本。该数据集包含大尺度多目标图像、小尺度对象图像、高相似度对象图像和密集分布对象	https://doi.org/10.1016/j.compag.2020.105585
12	西红柿害虫数据集[27]	8	互联网中收集到的8种常见的害虫，原始图片609张，数据增强后共4263张	https://data.mendeley.com/datasets/s62zm6djd2/1
13	桔小实蝇等六种常见果园害虫图像数据集[28]	6	包含桔小实蝇、金龟子、梨小食心虫、青叶蝉、星天牛、柑桔大实蝇6个种类。原始图像1613张，具有显著性特征图片799张，增强后2412张样本	https://www.agridata.cn/data.html#/datadetail? id=286640

2.2 部分数据集中样本分布

由表1可以看到，病虫害数据集大多针对特定的实际需求而建立，种类较少、数据量较小的大多为私有数据集。农业病虫害数据集不同于常见的深度学习数据集，该数据集中的一些样本对象为生活中不常见的样本，样本对象往往较难寻找和采集。这些客观原因导致了数据集类内容量缺乏、类间缺乏多样性、类别不均衡等问题较突出，同时病虫害图像本身还具有目标小、遮挡和一张图像中目标对象密集分布等特点[29]。

对部分开源的农业病虫害图像数据集的数据分布进行分析。参考相关统计量，选取了母体标准差、偏度系数、峰态系数三个统计量以及分辨率和标注信息进行样本分析，结果见表2。标准差可以反应数据的离散程度，由于表2中使用了数据集中的全部数据，所以计算的是母体标准差，偏度系数反应了数据的对称性，当其大于0时表明数据呈右偏，小于0时数据呈左偏，等于0时数据呈正态分布；峰态系数反应了一组数据峰值高低的特征，当其等于0时，表示数据接近于正态分布，峰态系数越低表明数据分布越平坦。根据美国电影电视工程师协会指定的高等级高清数字电视格式标准，可将图像分辨率指标分为：1080P以上的大分辨率，720—1080P的中分辨率，以及720P以下的小分辨率[30]。

由表2可以看到，在公开数据集中，母体标准差相对较大，尤其是样本容量和类别数较多的数据

表2 部分病虫害公开数据集分析结果

Table 2 Analysis results of some public datasets of diseases and pests

序号	数据集名称	类别数	样本容量/张	母体标准差	偏度系数	峰态系数	分辨率类型	标注信息
1	IP102	102	75,222	966.59	3.73	14.45	小	有
2	Pest	7	4639	357.04	0.70	−0.89	小	无
3	Plant Village	38	54,303	1158.27	2.73	5.70	小	无
4	西红柿虫害	8	609	32.52	0.05	−0.94	小	无
5	果园害虫	6	1568	169.27	−0.02	−1.50	小	无
6	Rice Leaf Disease Images	4	5932	118.70	−0.72	−1.44	小	无
7	苹果叶片病害	5	24,348	374.38	−0.20	−1.73	小	无
8	Wheat Leaf Dataset	3	407	51.18	1.71	−1.49	大	无

集有着更大的母体标准差，表明该数据集不同类别间样本容量离散程度较大，不同类别间的样本数量有较大的差距；由偏度系数可以看出，大多数据集存在长尾的现象，类内样本容量的均值在峰值的右边，呈不对称分布；从数据的峰态系数可以看到，IP102和Plant Village数据集的峰值更加集中，个别类别中样本容量较多，其他数据集的峰值呈平顶峰分布较为分散，类间的样本容量分布也较平缓；图像分辨率也是影响算法性能的重要因素之一，通过分析分辨率可以看到，较多的公开数据集中图像的分辨率相对较低，较低的分辨率会带来较小的模型计算量（FLOPs），减轻计算负担，但同时也会降低算法精度[31]。由这些因素演化而来的问题都会对算法性能产生影响，因此，总结农业病虫害图像数据集构建过程中的问题并分析问题产生的原因将有利于算法性能的提升。

2.3 构建农业病虫害数据集面临的问题与挑战

在农业病虫害识别和监测预警技术的发展过程中，农业病虫害数据集起着至关重要的作用。农业病虫害数据集经历了从单一病虫害、单一作物到多病虫害、多作物的发展历程，样本数量也从几百幅图像到上万张图像。然而在病虫害数据集的构建过程中，存在一些问题导致该领域缺乏高质量的数据集，影响了农业病虫害识别技术的发展和应用。根据表1和表2中常见的数据集以及类间样本分布情况，农业病虫害数据集构建所面临的问题与挑战总结为三个层次：分别是数据集层次，数据样本层次和使用层次，如图1所示。

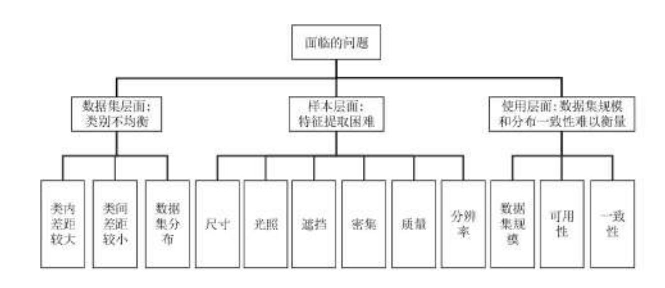


图1 农业病虫害图像数据集构建面临的挑战

Fig.1 Challenges in constructing agricultural pest and disease image datasets

2.3.1 数据集层面上类别不均衡

数据集层次中的类别不均衡表现在三个方面，类内差距、类间差距和数据集分布一致性。类内差距较大，指的是属于同一类别的图像具有较大差距的外表特征，如图2（a）所示，同样种类的害虫，却有着不一样的外观特征。类间差距较小，指的是属于不同类别的图像具有较小的外表特征，如图2（b）所示，不同种类的害虫有着相似的外观。因为部分害虫虽然外观相似，但是其足、雌雄外生殖器、卵和幼虫形态、化石形态等差异较大导致其属于不同的类别。

数据集分布指的是不同类别内的害虫，图像样本的数量以及相同样本的信息量分布一致。类别不均衡和同样的样本数量下携带的信息量不均衡都称之为不均衡数据集[32]。图像识别深度学习算法需要大量的高质量数据的支持[33]，农业病虫害数据



(a)较大的类内差距



(b)较小的类间差距

图2　不合理数据的类间和类内差距

Fig.2 Inter class and intra class differences of unreasonable date

集受到客观条件的影响，高质量的样本十分缺乏。一方面是受到自然环境的影响，有一些农业病虫害原始图像数据的获取十分困难，另一方面是对同一病虫害个体采集的数据过多，导致原始图像数据过于相似，如图3所示。

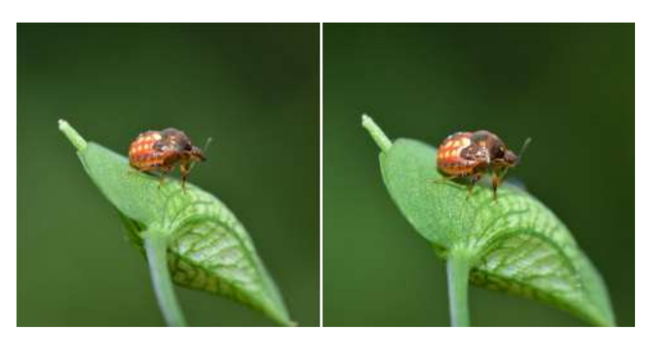


图3　整体相似的图像

Fig. 3 Overall similar images

数据集应当满足不同类别内的数据量分布一致，较小的类间差距、较大的类内差距以及不均衡的数据集都属于数据集层次的问题，会对算法的训练结果带来一些较为严重的影响[34]。

（1）过拟合。当数据集中的样本图像数量较小时，模型容易产生过拟合的现象，如果数据集中的大多数样本较为集中在某几个类别中时，深度的训练导致模型过拟合，模型会出现在含有样本容量较少的类中表现较差，模型的鲁棒性和泛化能力较差。

（2）域偏移。指在大规模训练集上训练的模型在应用于具有不同统计量的目标数据集时表现不好[35]，当源数据样本较少时，模型往往会根据数据基类中的大规模数据来提取通用特征。当目标数据集中的样本较少时，源数据集往往会与目标数据集存在较大的差异，两个数据集之间公共的特征较少。

（3）数据分布较差。当数据集中图像样本的数量偏少时，会导致数据偏差和分布偏差的问题。较少的训练样本在一定程度上会放大噪声的影响，可能会使类内样本间的距离偏大，而类间的图像样本距离偏小，同时较小的样本数量使得模型无法准确、完整的表示样本数据的真实分布，目标对象与背景相互影响，从而降低模型的准确率。为了解决不均衡数据集对模型性能的影响，有专家提出使用欧拉距离、交集距离和二次方距离度量图像间的颜色特征（Hue，Staturation，Value，HSV）和纹理特征（Local Binary Pattern，LBP）直方图，进行相似性判断，过滤掉相似度较高的图像[36]。He等[37]采用几何增强的方式增加图像数据的数量，包括翻转、裁减、缩放、变形等，达到数据集种类平衡。Chodey和Noorullah Shariff[38]采用了强度增强的方式，包括指数变换、对数变换、线性变换等方式扩充数据。范馨月等[39]对长尾数据集采用基于目标尺度的方法进行数据增强，增加小样本的数量，对其进行重采样。部分学者[40-43]通过增加数据集中小目标和密集样本的数量和改进识别算法，增强了提取小目标和密集样本特征的能力。

2.3.2　样本层面上样本特征提取困难

受到农业实际应用场景及拍摄者主观因素的影响，采集到的图像中的目标对象往往具有目标过小、目标过大、目标密集、目标间遮挡、图像部分模糊和分辨率过大或过小等特点。如图4所示。

数据集中目标在图像中的尺度变化范围较大，给检测和识别带来了许多难点。Li等[44]通过图像中的目标与图像比例来衡量图像的复杂度，认为比例越小，图像越复杂，其包含的本身特征较少并且包含的背景噪声较多，容易受到噪声的影响。在模



(a)小目标对象 (b)图像部分模糊 (c)目标密集图像 (d)目标间遮挡图像 (e)大目标图像 (f)不同光照强度下的目标图像

图4 在样本层次上不同特征的图像

Fig. 4 Images with different features at the sample level

型的卷积层提取特征信息时，随着层次的加深，会导致特征信息的丢失[45]，中浅层神经网络能够较好地提取到小目标的特征，而大目标的特征需要深层次的神经网络模型进行提取。但是大目标与小目标间的差距过大，在神经网络加深的过程中可能导致模型对图像中的小目标出现漏检的现象[46]。密集和遮挡的目标对象会导致特征提取不完整，遮挡较多的目标也可能出现漏检的现象，同时被遮挡物分割为多段的目标也难以判断是否属于同一目标[47]。

图像中目标模糊也是影响图像质量的主要因素。在拍摄过程中，使用微距镜头易受到人为或者景深的影响导致图像出现部分模糊，或者全部模糊的情况。模糊的部分可能导致模型提取到目标的部分特征丢失，如图4（b）所示。不同的光照和不同的图像分辨率对目标检测和识别也有着较大的影响。如图4（f）所示，左边为强光下的图像，右边为自然光较弱的环境下拍摄的图像，其图像特征有着明显的不同。从图4（f）可以看到强光环境下，害虫的色彩饱和度增加了，背部的细节轮廓也较为明显，有利于算法对特征的提取。而弱光的环境下，图像整体偏暗，背部的细节特征较难分辨，甚至部分区域的颜色区分度不高，影响算法对特征的提取。

在病虫害图像识别任务中，病虫害的背景也有一定的规律。在自然环境下的病虫害图像背景往往是作物的茎叶和土地，而实验环境中拍摄到的图像背景往往过于单一，相对于实验环境中的背景目标，大田环境中的图像识别难度更高，如图5所示。图像中的背景单一或者目标尺寸过大可理解为图像复杂度过低，Borji等[48]通过计算图像的熵来衡量一幅图像的复杂度，熵值高的图像通常由更多的物体和更多的纹理构成，熵值越高，图像复杂度越高。





(a)单一背景



(b)大田背景

图5 不同背景的害虫样本图像

Fig. 5 Pest sample images with different backgrounds

图像的分辨率对算法识别结果的影响也不可忽略，较高的图像分辨率虽然需要更多的内存空间来进行训练，但是其包含的特征信息也更多。而较低的图像分辨率会导致模型在下采样时丢失很多特征信息，影响算法的准确率。目前，大部分算法在提取特征前会对图像进行resize处理，将图像不同的分辨率缩小或者增大到一个固定的值。在缩放操作时可能会导致目标出现模糊的情况，也会影响算法的识别结果。

2.3.3 使用层面上数据集规模和分布一致性难以衡量

数据集使用层面的挑战来自于数据集的规模、可用性、训练样本和测试样本分布一致性等问题。计算机视觉的迅猛发展离不开大规模标签数据的产生。Sun等[49]通过实验证实，在3亿张样本图像的情况下，抑制单个样本的噪声和扩大与数据集匹配的模型深度后，随着数据集规模的增大，视觉任务模型的性能也随之得到提升。计算机视觉任务常用的MSCOCO公共数据集具有80个类别，分别包含118,287个训练样本、5000个验证样本和40,670个测试样本。从表1可以看到与MSCOCO数据集相比，农业病虫害图像数据集规模还有待进一步扩大。在农业病虫害识别检测任务中，数据集的可用性往往是首先考虑的因素，具体指的是数据集是否方便地获取和使用。部分农业病虫害图像数据集属于私有数据集，不方便获取，给相关检测识别任务带来了获取难度；部分公开的数据集中图像的标注准确性较差，含有噪声较多，也增加了相关检测识别任务的使用难度。

模型评估的默认假设前提是训练数据样本和测试数据样本的分布形式具有一致性，研究者应当重视该领域中的数据分布一致性问题，但实际应用中却往往容易忽视。训练数据和测试数据的分布一致性是指在度量方法下度量的训练和测试数据分布的一致性，主要由于选择偏差引起，农业病虫害图像数据采集的难度导致了在数据集的构建过程中容易在训练样本和测试样本上出现分布差异，这种差异主要体现在选择偏差上[50]。数据偏差可能导致识别模型的泛化能力下降、模型中的有偏估计等，因此训练和测试数据分布的一致性至关重要。图像中的偏差问题很大程度上已经影响了图像识别算法的实际使用效果[51]，有专家认为数据集的偏差主要来自于原始图片的采集与标注，将采集与标注分开进行可以避免设计上的偏差[52]。Bylinskii等[53]提出要充分理解和利用现有带有偏差的数据集，并且根据实际项目任务的需要，去处理和构建基准数据集。数据集中图像样本的选择偏差和复杂度是构建数据集时要考虑的两个方面[54]，需要通过定量进行分析，Borji等[55]通过平均主视图（Average Annotation Map，AAM）的方法来进行评价，它将数据集中所有的标注图用伪彩色图进行表示，颜色较深越靠近图像中央的区域表明该图像的选择偏差越大。Fan等[56]使用目标轮廓中心到图像中心的距离归一化（Normalized Object Distance from Image Center，NOD）来度量中心偏差，该归一化距离等于目标轮廓的中心到图像中心的距离除以整张图像对角线长度的一半，距离越小表明图像选择偏差越严重。

在实验环境中采集到的训练数据集中，农业害虫样本图像目标显著性较强、图像背景较为简单、光线单一，而实际测试使用环境中的图像却存在背景复杂、光线多变等因素。这些偏差都会导致训练和测试数据集分布的差异，从而影响模型的使用性能。

3 数据采集环境与方法

构建农业病虫害数据集所面临的问题与其特殊环境和采集方式相关。下面分别从数据采集环境和数据采集方式两个方面来分析构建数据集所面临的问题。在农业病虫害识别应用中，害虫图像的采集是十分重要的环节之一，同时也是非常耗时的环节。在不同环境下利用不同的数据采集方法，所构建的数据集也具有其独到的特点。

3.1 数据采集环境

当前大多数据集是在相对理想环境的实验室中利用专业设备采集到的图像数据，这种方式能够快速获得想要的特定样本图像，种类往往不具有足够的代表性[57]。在实际农业生产中，不同地区、不同作物的病虫害图像数据采集是非常困难的。第一，随着农药的普及和及时使用，大田作物中的一些病虫害更难以进行全周期采集，想要采集到完整的图像数据越发困难。第二，在自然环境中的农业病虫害，发生规律差异很大，时间横跨一年四季、地理位置遍布祖国大江南北，甚至同种病虫害不同发展阶段的形态也差异明显，准确地鉴别并对其采集也是一项严峻的挑战；第三，农业病虫害由于个体小、隐蔽性强，田间实地难以发现，因此发现并采集到清晰的图像样本也是十分困难的。表3对比

了农业病虫害图像不同采集环境的特点。

表3 病虫害图像的不同采集环境对比

Table 3 Comparison of different collection environments for disease and pest images

环境	特点
实验环境	周围环境可控、病虫害种类较少、便于采集图像
大田自然环境	周围环境多变、病虫害受季节、温湿度等影响明显
自然光源	光线不可控、受位置、季节、时间等影响较大
人工光源	光线颜色和强弱相对可控、利于拍摄细节特征

周围环境对农业病虫害图像采集有着较大的影响，大田自然环境中，寄主、季节、温湿度、光线等环境因素的多变导致了农业病虫害所呈现出的性状具有多样性。图6展示了部分害虫在不同环境中呈现的外观多样性。图6（a）为大田自然环境中同一种害虫，受到寄主、温湿度、光线角度等因素影响，表现出的背部不同颜色特征。图6（b）为同一种害虫在不同光源条件下的外观特征，在自然环境中光线较弱或者受到遮挡时，图像整体偏暗，会出现目标的部分特征消失现象，在人工补光适当的条件下，图像中的目标和背景表现出的特征较为明显，当人工补光过度时，光线整体较亮，会出现部分背景的特征消失现象。图6（c）展示的是受到不同季节、不同寄主等因素的影响，同一害虫表现出的不同颜色对比。通常而言，在多样的环境和不同光线下的复杂场景中采集到的病虫害图像有利于丰富数据集的多样性，更有利于提高模型的鲁棒性。因此应当根据实际任务需求，灵活地调节环境变量因素，构建数据集。



(a)不同光源角度下采集的图像



(b)不同光源下采集的图像



(c)不同自然环境下采集的图像

图6 不同环境下拍摄的害虫图像

Fig. 6 Images of pests taken in different environments

3.2 数据采集方法及设备

目前越来越多的团队根据自己的需求，使用不同的设备来采集特定种类的农业病虫害数据。早在1998年，Zayas和Flinn[58]就通过照相机等设备在室内实验室环境中搭建了图像采集设施，利用不同角度的灯光照射，解决了拍摄图像中存在阴影的问题。韩瑞珍[59]开发了一种害虫样本图像采集系统，该系统利用工业相机，将害虫样本置于置物台上，分别进行正视和俯视两种角度的拍摄。刘媛媛[60]为了拍摄清楚害虫的细节特征，使用了显微镜采集稻纵卷叶螟、小菜蛾等图像数据。Wu等[61]使用无人机搭载摄像头对健康和患病的玉米数据图像进行采集。周瑶[62]在大田环境安装害虫引诱设备来捕获害虫，通过在引诱设备对面安装摄像机来采集害虫图像。白静亚[63]开发了一款带有自动采集害虫图像的诱虫灯设备，利用害虫屈光原理，引诱害虫撞击昏迷，在害虫以自由落体的过程中对其进行拍摄。随着手机摄影技术的发展，智能手机的拍照摄像功能也越来越强大，使用手机采集农作物病虫害图像数据更加便捷，Li等[64]通过智能手机和索尼单反相机采集了像素大小为4928×3264的水稻鞘叶枯病图像1800张、水稻褐斑病图像1760张，以及像素大小为2592×1944的水稻干螟虫症状图像1760张。

近年来，许多团队也开始尝试将病虫害图像数据和其周围环境数据进行融合，如地理信息、气象信息、作物信息等[65]，形成多模态农业病虫害数

据集，用以挖掘更具有价值的信息。史东旭等[66]使用最新的通信技术，通过无人机搭载摄像头和传感器来采集图像信息和周围环境信息，利用大数据平台建立农业病虫害发生模型，从周围环境信息来分析农业病虫害的发生原因。为了进一步分析不同数据采集方式带来的影响，表4整理了不同设备采集农业病虫害图像数据方式的优缺点。

表4 图像的不同采集方式对比

Table 4 Comparison of different image acquisition methods

序号	设备类型	优点	缺点	适用场景
1	智能手机	灵活方便、简单易操作	图像清晰度不够	室内、户外
2	单反相机	灵活方便、图像清晰、能凸显较多的细节	需要掌握较专业的拍摄技术	室内、户外
3	无人机平台	能更多地拍摄病虫害群体特征	拍摄不到病虫害个体特征	户外
4	自动采集装备	自动化操作、效率高	价格昂贵、一旦安装完毕，只能拍摄特定的种类	室内、户外
5	显微镜	更多凸显图像细节特征	携带不方便、拍摄过程较为麻烦、拍摄背景单一	室内

目前，农业病虫害图像数据采集设备可分为3大类：手持设备、无人机平台和固定采集装置。手持设备具备方便灵活等优势，将网络设备、单反相机和拍摄杆等设备组合到一起能够取长补短，拍摄到更多种类的病虫害图像，这类设备便于在自然环境复杂的地方拍摄[67]，但是往往需要具有植保知识的专业人员去大田环境或者实验室里找到病虫害的发生位置，在拍摄过程中还需要具备一定的摄影知识，拍摄效率低；固定采集装置可安装在室外大田环境或者室内实验室中，一旦固定装置建造好，往往只能拍摄固定区域内的固定种类的害虫图像，造价成本比较昂贵，但是固定装置可搭载高精度的监测设备全天候不间断自动拍摄，并且将获取到的图像数据及时上传至云平台进行处理[68]，拍摄效率较高；无人机平台可以灵活搭载各种光学传感设备，例如高清摄像头、高光谱镜头和远红外镜头等[69]，其更适用于在大型农场或者大规模连片大田进行图像数据采集，但是其拍摄的细节特征不够明显，多用来进行农业病害图像采集，少有学者用它来采集农业害虫图像。在不同的环境中，灵活应用不同的设备可以提高图像采集效率，但是不同的设备所采集到的图像也具有不同的特点。

智能手机受到硬件的限制，拍摄出的照片清晰度往往不高，单反相机属于专业拍照设备，微距镜头的使用容易导致拍摄出的照片出现模糊或者光线较弱的现象，大田自动采集设备需要事先设定好的拍照程序，在一定时间内易出现遮挡、密集等问题，显微镜可以拍摄尺寸更小的目标，受到其视野的限制，照片中的背景信息较为缺乏。一套便携、优价、自动化程度高、适用场景广、能够采集多源数据的农业信息化采集设备可以帮助农业科技人员提高数据采集效率，应当加强农业设备和农业信息化的融合，为农业病虫害识别提供强大的基础采集设备，为病虫害监测预警提供有力的数据支撑。

4 农业病虫害图像数据标注

除了图像采集的环境和方法，图像标注质量也是影响数据集质量的重要因素。图像识别的前提是要通过训练数据来告诉计算机一幅图像中真实的样本对象，再通过算法从这些大量的样本对象中提取出属于某一种类别的特征。

4.1 图像标注任务

数据标注的目的就是要将算法要识别的图像提前打上标签，计算机在这些打上标签的目标对象中提取目标特征，最终实现计算机自动识别目标对象[70]。数据标注尚无统一的定义，Zhu等[71]将数据标注定义为对未处理的原始数据进行加工处理转换为计算机可以识别的过程，标注对象可以是图像、视频、文本、语音等。对于图像类型的数据，常见的标注格式主要有矩形框标注、多边形框标注、描点标注和分类标注等[72]，在农业病虫害图像识别领域中，最常用的标注格式是矩形框标注和多边形框标注。矩形框标注常用的标注工具是La-

belImg，该工具操作便捷，会生成包含矩形框位置信息的XML格式文件，用户只需要拖动鼠标将待选中的目标用矩形框框出即可，如图7（a）所示，这种标注方法一般在进行分类或者检测识别时使用的较多；多边形框标注常用的标注工具是LabelMe，该工具可以标注图像和视频，可以导出VOC和COCO数据格式，用户需要将待选目标的轮廓用多边形勾勒出来[73]，如图7（b）所示，这种标注方法一般在进行图像语义分割或者实例分割时使用较多。



(a)矩形标注　(b)多边形标注

图7 不同目标标注方式

Fig. 7 Different target annotation methods

4.2 图像标注流程

农业病虫害图像的标注需要较强的专业背景，目前大多以人工标注为主。在标注过程中可以将标注者分为3种不同的角色[74]。

（1）标注员。负责具体图像的标注工作，需要具备一定的农业病虫害相关背景知识，大多数团队为了提高标注效率，多让具有农业背景的高校学生进行标注。

（2）审核员。负责对标注员已经标注好的数据进行审核和校验，对已经标注好的图像数据进行标注质量检验，分别判断其标注是否完整、标注对象与标签是否吻合，对于标注有误的图像进行修改或者打回重标，审核员需要具备较强的农业病虫害背景，能够识别辨别各种病虫害图像，因此审核员往往是农业植保领域的专家来担任。

（3）管理员。负责对原始图像以及标注后的图像进行分类管理和统计，最终形成可使用的农业病虫害图像专业数据集。

不同角色的标注工作缺一不可，他们之间的有机结合，可以在一定程度上保证图像标注质量。随着需要标注的数据不断增加，为了提高标注效率，在人工智能领域中，大多数的数据标注采用众包方式完成，流程如图8所示。

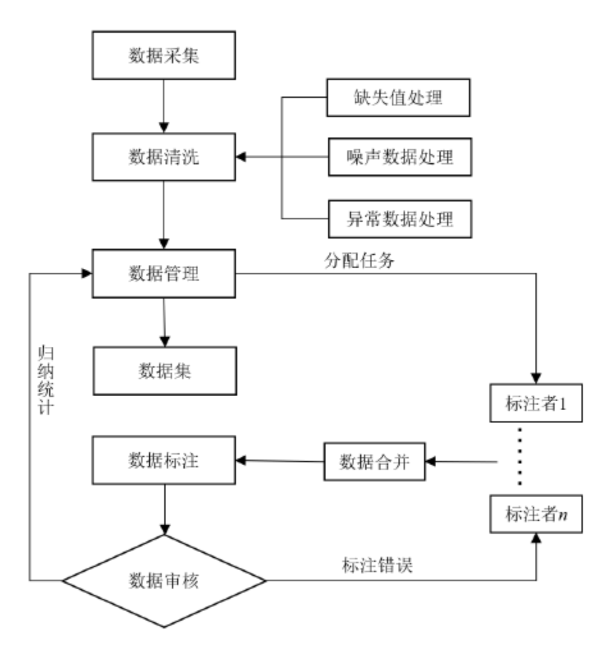


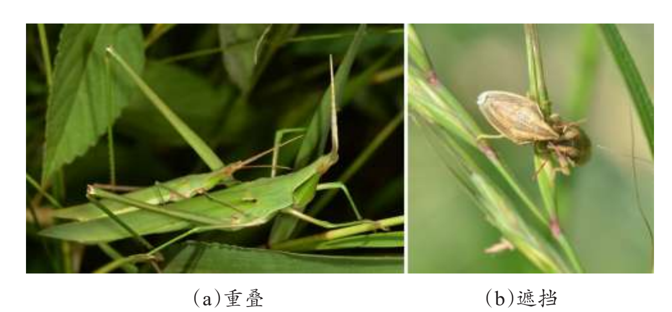
图8 农业害虫图像数据标注流程

Fig. 8 Annotation process of image data

图像数据标注流程从数据采集开始，采集的方式可以为室内实验环境或者室外自然环境，采集的工具有单反相机、智能手机、固定采集装置和无人机平台等，通过这些设备采集到的图像数据为原始数据，可能存在缺失、噪声、异常等问题，需要进行第二步清洗[75]。清洗后的数据将交给数据管理员，由管理员对这些原始数据进行统计分类和任务分配，将待标注的数据按照一定规则进行切片分发给不同的标注者单独标注。当数据按照标准规范被标注好后，再对这些数据按照规则进行合并处理，分发给审核员进行审核，负责将符合满足标注质量的图像数据转发给数据管理员，由数据管理员对这些图像数据进行归纳统计，形成可直接使用的数据集。

图像的鉴别和标注会影响模型的准确性，鉴别错误的图像会被打上错误的标签，在训练后也会得到错误的结果，降低算法的准确性。同样，不准确标注的图像也会使算法降低准确性。在标注过程中经常会出现一些难以标注的目标，如图9所示，这类图像经常存在于自然界中。对这些图像的标注比

较困难，标注不当会引入较多噪声，会对模型识别结果产生较大影响。



(a)重叠 (b)遮挡

图9 难以标注的复杂图像样例

Fig. 9 Complex images that are difficult to annotate

5 数据集质量评价方法

目前对于高质量的农业病虫害数据集评价指标尚无统一的定义，但是可以肯定的是，构建高质量的数据集对于模型的性能发挥着重要的作用。针对构建农业病虫害图像数据集面临的一些问题和挑战，本文从数据分布一致性、数据集规模和数据标注质量三个方面总结了现有的相关评价方法。

5.1 数据分布一致性评价

判断训练数据集和测试数据分布是否一致的度量方法常分为度量函数方法和假设检验类方法[76]。假设检验类方法是衡量样本与样本，或者样本与总体之间差异性的一种方法，事先通过对训练和测试数据的分布进行假设，然后利用检验统计量对数据分布进行一致性检验。度量函数的方法因其简单直观，被多数文献采用。度量距离常见的有Hellinger距离、全变差距离和相对熵距离（Kullback-Leibler，KL）等[77]，如公式（1）是用f散度度量函数计算两个分布间的距离

$$D(P(x),Q(x))=\int f(\frac{dP}{dQ})dQ \tag{1}$$

其中，$P(x)$和$Q(x)$分别为两个数据分布的密度函数，当$f(x)=\left(\sqrt{x}-1\right)^2$时，该度量函数度量的是Hellinger距离；当$f(x)=\frac{1}{2}|x-1|$时，该度量函数度量的是全变差距离；当$f(x)=x\log x$时，该度量函数度量的是KL距离。

当训练数据和测试数据存在误差时，可能导致算法性能下降的问题，因此需要对训练数据和测试数据进行校正。目前常用样本自适应分布差异校正和特征自适应分布差异校正两种方法，前者采用相关机器学习算法对模型测试效果较差的数据进行训练，还有专家利用测试集和训练集中较为重要的数据对损失进行加权对数据分布进行校正[78]。后者经常将测试集和训练集中的数据特征进行转换，但保留数据原有的特征结构，利用新特征来代替旧特征，形成对应关系[79]。

5.2 数据集规模评价

数据集规模对于模型的训练有着重要作用，模型的准确性和泛化能力与数据集的规模有着高度的相关性。公式（2）直观地表现了几个因素之间的关系[80]。

$$loss=(bias)^2+variance+noise \tag{2}$$

其中，$loss$表示损失函数；$bias$表示模型的偏差，是真实标签与预测标签之间的偏离程度，刻画了模型的拟合能力；$variance$为模型的方差，刻画了模型的稳定性；$noise$是模型的噪声，表示当前模型所能达到的期望误差下限。

当数据量一定时，模型必须在方差和偏置之间进行权衡，根据经验法则，数据集的容量应当是提取特征数据的十倍。要使模型的泛化能力增强，其预测精度就会下降，反之提升模型的精度，其泛化能力就会下降。在保证标注噪声非常小的前提下，解决这一问题的最好方法就是提升数据集规模。

Sun等[49]选择了分别包含128万张图像、1400万张图像和3亿张图像的不同数据集探究计算机视觉模型的容量和数据集规模之间的关系，实验结果表明，增加数据规模的同时，扩大模型的容量，模型的准确性会得到不断提升，模型的性能表现随着数据规模的增大呈对数关系提升。

当数据量一定时，对于深度学习模型而言，选择与模型的深度相匹配的数据集规模很重要。当前根据比较著名的VC维（Vapnik-Chevronenkis Dimension）来评估模型需要的训练数据规模[81]，训练数据的规模与VC维之间存在一个特定的函数关系，如公式（3）所示。

$$N = F\left(\frac{VC + \ln\left(\frac{1}{d}\right)}{epsilon}\right) \tag{3}$$

其中，N为与模型匹配的数据集规模；利用函数F（x）对模型进行计算可以得到与之匹配的数据集规模；VC维表示模型的复杂程度，模型复杂度越高，VC维也越大；d是模型的错误率；*epslion*是模型的误差率。可以看到，利用VC维对数据集规模进行评价的过程中，数据集规模与模型的复杂程度相关。

5.3 图像标注质量评价

图像标注质量的好坏与特征提取是否全面相关。农业病虫害图像识别算法是根据图像中的像素点进行训练的，因此标注者能否准确地判定像素点关系到整张图像标注的质量优劣[82]，标注的像素越接近于物体的真实边缘，标注的难度越大，标注的质量相应越高，反之标注质量越差。想要保证标注的准确性达到100%，则标注的范围应当与将要识别的真实物体边缘相差不超过1个像素。Gupta等[83]定义了像素精度误差（Pixel-wise Accuracy）ε来计算像素级标注质量，如公式（4）所示。

$$\varepsilon = \frac{1}{W \times H}\sum_{x=1}^{W}\sum_{y=1}^{H}\left\| M(x,y) - G(x,y) \right\| \tag{4}$$

其中，M（x，y）为标注后的区域像素；G（x，y）为图像中真实的目标区域对应的像素；W和H分别为图像的宽度和高度，通过公式（4）计算出M和G的像素精度误差ε。当ε越接近于1，表明标注像素和图像目标实际像素相差越大，标注质量较差；当ε越接近于0，表明标注质量越好。

标签的正确性也是图像标注的质量好坏的重要方面[84]，常用的评价指标主要有多数投票算法（Majority Voting，MV）、期望最大值算法（Expectation Maximization，EM）以及RY算法。MV算法的主要策略是选择大多数标注者都认为正确的结果[85]，如公式（5）所示。

$$\widehat{y_i} = \begin{cases} 1, & \frac{1}{M}\sum_{j=1}^{M} y_i^j > \frac{1}{2} \\ \text{random}, & \frac{1}{M}\sum_{j=1}^{M} y_i^j = \frac{1}{2} \\ 0, & \frac{1}{M}\sum_{j=1}^{M} y_i^j < \frac{1}{2} \end{cases} \tag{5}$$

其中，y_i^j表示标注者对样本图像的预测标签，例如在数据样本中有待标注的图像有m个，每一个图像都对应着一个二元分类，将这些图像样本通过众包分配给M个标注者进行标注，则每个标注者j都会对图像i作出预测，得到y_i^j预测值，最终得到该图像的所有标签为$\{y_i^1, y_i^2, \cdots\cdots, y_i^M\}$，然后根据这些预测值选择超过一半以上的标注者认为是正确的标签作为最终标签，但是该算法没有考虑到单独标注者的可靠性。

事实上，大多数人工作出的选择不一定是正确的，基于此，Raykar等[86]提出了一种使用最大期望值的EM算法，该算法提出在利用标注者标注错误的数据构建错误率混淆矩阵，与实际观测的结果进行比较，当比较后的差异较小时，就说明该标注的质量越高。任何一个标注者对目标对象的标注可以看作是一个二分类问题，即标注正确与标注错误。在二分类问题中常用精准率（Precision）、召回率（Recall）和F_1分数这三种指标来构建混淆矩阵[87]，如表5所示。TP为真实的正样本被预测为正样本数量，FN为真实正样本被预测为负样本数量，FP为真实的负样本被预测正负样本数量，TN为真实的负样本被预测为负样本数量。

表5 二分类混淆矩阵

Table 5 Binary confusion matrix

预测值/实际值	实际正样本	实际负样本
预测正样本	TP	FP
预测负样本	FN	TN

精准率定义为在所有预测结果的正样本中，真样本数量所占的比重，如公式（6）所示。

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

精确率只能反应当预测结果为正样本时的可靠程度。但其存在的问题是。当在结果中仅仅有一个正样本被预测为正样本时，即只有一个标注者将正确的目标对象标注为正确的类别，该模型的精确率为100%。

召回率定义为在所有的预测结果中预测正确的正样本数量与所有实际为正样本的数量比例，如公式（7）所示。

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

当结果中所有的样本全部被预测为正样本时，即标注者将所有的目标对象都标注为同一个类别时，召回率为100%，因为召回率仅仅关注正样本的情况。

F_1分数调和了召回率和精准率之间的缺点，综合了两者的结果，如公式（8）所示。

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (8)$$

F_1分数表示预测的结果中真实的正样本所占的比重和其是否可靠。当召回率和精准率都比较高时，F_1分数的结果才会高，当结果越高时，表明标注正确性越高。

Raykar等[88]和于洪等[89]结合了MV和EM算法特点，提出了RY算法，该算法利用公式（2）提出了用于表示标注者本身特征的敏感性（specificity）和特异性（sensitivity）概念[90, 91]，通过对标注者的敏感性和特异性进行分析，排除了不合格标注者标注的图像，从而提高了标注质量。

6 数据集构建建议

本文系统地综述了构建农业病虫害数据集的现状及存在问题的原因，目前病虫害识别领域中缺乏大规模、高质量的数据集。本文从农业病虫害图像数据本身存在的问题入手，总结了构建农业病虫害数据集主要面临的挑战，从图像数据集的采集和标注两个关键环节分析了形成这些挑战的原因，总结了相关数据集质量的评价方法，提出以下3点数据集构建建议。

（1）结合实际使用场景构建农业病虫害数据集。在构建数据集时，应当充分考虑算法的使用场景。在实际生产环境中。使用者往往给出的将要识别的图像更多来自于田间地头随手拍摄的农业病虫害照片，由于病虫害的个体小，拍摄出图像中的目标往往更小，显著性更低。在拍摄图像时，应当将目标显著性图像与目标非显著性图像置于同等地位[92]，同时增加模糊无语义的图像和不同角度、不同光线下的图像。合理划分数据集中的类别，多视角、多环境下拍摄图像，可从算法提取特征进行分类的角度，满足较大的类间距离和较小的类间距离，保证各类别中的数据和特征分布尽可能保证均匀。数据集图像数据中心中的目标对象不应存在选择偏差，应当包含多种目标对象位置，多种简单、复杂的背景[93]，中小型目标对象的图像样本尽可能多，为农业病虫害识别提供高质量的数据支持。

（2）平衡数据集与算法间的关系。数据集的规模、单张图像标注的准确性以及数据采集方法都会影响到模型的性能。应当结合使用场景，探究农业病虫害图像数据的规模与模型性能之间的关系，挖掘数据规模与算法性能的平衡点，为构建病虫害数据集提供规模依据。规范图像标注的方法，在遮挡、模糊、密集等场景下，探究不同目标对象标注方法与模型性能之间的关系，尽可能减少标注过程带来的噪声，提高模型的性能。

（3）增强数据集的使用价值。农业病虫害数据集建设是一项长期坚持的工作。为了扩大数据集的价值、丰富其使用场景，大规模基础病虫害数据集不应当只有图像数据，应该还包含文字、视频、图像周围环境信息等多模态数据。为了适应较快的业务创新速度，应构建与整合多模态农业病虫害数据资源，建设农业病虫害大数据中台，将业务逻辑中的数据存储和计算力抽离，由数据中台对海量数据进行计算、存储、加工和统一标准，为各业务系统和具体的落地项目提供高效服务，简化业务系统的复杂性，让研究者更专注于应用模型研发。

利益冲突声明： 本研究不存在研究者以及与公开研究成果有关的利益冲突。

参考文献：

[1] 雷仲仁, 郭予元, 李世访. 中国主要农作物有害生物名录[M]. 北京: 中国农业科学技术出版社, 2014.
LEI Z R, GUO Y Y, LI S F. Catalogue of pests on major crops in China[M]. Beijing: China Agricultural Science and Technology Press, 2014.

[2] 吴钜文, 陈红印. 蔬菜害虫及其天敌昆虫名录[M]. 北京: 中国农业科学技术出版社, 2013.
WU J W, CHEN H Y. Catalog of insect pests and their natural enemies of vegetable crops[M]. Beijing: Agricultural Science and Technology Press, 2013.

[3] 翟肇裕, 曹益飞, 徐焕良, 等. 农作物病虫害识别关键技术研究综述[J]. 农业机械学报, 2021, 52(7): 1-18.
ZHAI Z Y, CAO Y F, XU H L, et al. Review of key techniques for crop disease and pest detection[J]. Transactions of the Chinese society for agricultural machinery, 2021, 52 (7): 1-18.

[4] HUO M Y, TAN J. Pattern recognition and artificial intelli-

gence[M]. Berlin: Springer, 2020, 404-415.

[5] 翁杨, 曾睿, 吴陈铭, 等. 基于深度学习的农业植物表型研究综述[J]. 中国科学: 生命科学, 2019, 49(6): 698-716.
WENG Y, ZENG R, WU C M, et al. A survey on deep-learning-based plant phenotype research in agriculture[J]. Scientia Sinica (vitae), 2019, 49(6): 698-716.

[6] 杭立, 车进, 宋培源, 等. 基于机器学习和图像处理技术的病虫害预测[J]. 西南大学学报(自然科学版), 2020, 42(1): 134-141.
HANG L, CHE J, SONG P Y, et al. Studies on pest prediction based on machine learning and image processing technologies[J]. Journal of southwest university (natural science edition), 2020, 42(1): 134-141.

[7] TANG Y C, CHEN C, LEITE A C, et al. Editorial: Precision control technology and application in agricultural pest and disease control[J]. Frontiers in plant science, 2023, 14: ID 1163839.

[8] 赵春江. 农业知识智能服务技术综述[J]. 智慧农业(中英文), 2023, 5(2): 126-148.
ZHAO Chunjiang. Agricultural knowledge intelligent service technology: A review[J]. Smart Agriculture, 2023, 5(2): 126-148.

[9] 黄凯奇, 任伟强, 谭铁牛. 图像物体分类与检测算法综述[J]. 计算机学报, 2014, 37(6): 1225-1240.
HUANG K Q, REN W Q, TAN T N. A review on image object classification and detection[J]. Chinese journal of computers, 2014, 37(6): 1225-1240.

[10] DE CESARO JÚNIOR T, RIEDER R. Automatic identification of insects from digital images: A survey[J]. Computers and electronics in agriculture, 2020, 178: ID 105784.

[11] LI W Y, ZHENG T F, YANG Z K, et al. Classification and detection of insects from field images using deep learning for smart pest management: A systematic review[J]. Ecological informatics, 2021, 66: ID 101460.

[12] 汪京京, 张武, 刘连忠, 等. 农作物病虫害图像识别技术的研究综述[J]. 计算机工程与科学, 2014, 36(7): 1363-1370.
WANG J J, ZHANG W, LIU L Z, et al. Summary of crop diseases and pests image recognition technology[J]. Computer engineering & science, 2014, 36(7): 1363-1370.

[13] HASAN R I, YUSUF S M, ALZUBAIDI L. Review of the state of the art of deep learning for plant diseases: A broad analysis and discussion[J]. Plants, 2020, 9(10): ID 1302.

[14] HORAK K, SABLATNIG R. Deep learning concepts and datasets for image recognition: Overview 2019[C]// Eleventh international conference on digital image processing (ICDIP 2019). Berlin, German: Springer, 2019: 484-491.

[15] ARSENOVIC M, KARANOVIC M, SLADOJEVIC S, et al. Solving current limitations of deep learning based approaches for plant disease detection[J]. Symmetry, 2019, 11(7): ID 939.

[16] CHOUHAN S S, SINGH U P, KAUL A, et al. A data repository of leaf images: Practice towards plant conservation with plant pathology[C]// 2019 4th International Conference on Information Systems and Computer Networks (ISCON). Piscataway, New Jersey, USA: IEEE, 2020: 700-707.

[17] DAVID P H, MARCEL S. An open access repository of images on plant health to enable the development of mobile disease diagnostics[EB/OL]. arXiv:1511.08060, 2015

[18] WU X P, ZHAN C, LAI Y K, et al. IP102: A large-scale benchmark dataset for insect pest recognition[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, New Jersey, USA: IEEE, 2020: 8779-8788.

[19] SETHY P K, BARPANDA N K, RATH A K, et al. Deep feature based rice leaf disease identification using support vector machine[J]. Computers and electronics in agriculture, 2020, 175: ID 105527.

[20] 陈雷, 袁媛. 大田作物病害识别研究图像数据集[J]. 中国科学数据, 2019, 4(4): 85-91.
CHEN L, YUAN Y. An image dataset for field crop disease identification[J]. China scientific data, 2019, 4(4): 85-91.

[21] 袁媛, 陈雷. IDADP-葡萄病害识别研究图像数据集[J]. 中国科学数据, 2022, 7(1): 86-90.
YUAN Y, CHEN L. An image dataset for IDADP-grape disease identification[J]. China scientific data, 2022, 7(1): 86-90.

[22] WANG R J, LIU L, XIE C J, et al. AgriPest: A large-scale domain-specific benchmark dataset for practical agricultural pest detection in the wild[J]. Sensors, 2021, 21(5): ID 1601.

[23] 周敏敏. 基于迁移学习的苹果叶面病害Android检测系统研究[D]. 杨凌: 西北农林科技大学, 2019.
ZHOU M M. Apple foliage diseases recognition in android system with transfer learning-based[D]. Yangling: Northwest A & F University, 2019.

[24] GEROVICHEV A, SADEH A, WINTER V, et al. High throughput data acquisition and deep learning for insect ecoinformatics[J]. Frontiers in ecology and evolution, 2021, 9: ID 600931.

[25] RUSTIA D J A, CHAO J J, CHIU L Y, et al. Automatic greenhouse insect pest detection and recognition based on a cascaded deep learning classification method[J]. Journal of applied entomology, 2021, 145(3): 206-222.

[26] WANG Q J, ZHANG S Y, DONG S F, et al. Pest24: A large-scale very small object data set of agricultural pests for multi-target detection[J]. Computers and electronics in agriculture, 2020, 175: ID 105585.

[27] HUANG M L, CHUANG T C. A database of eight common tomato pest images[DS/OL]. Mendeley Data, (2020-05-27) [2023-06-02]. https://doi. org/10.17632/s62zm6djd2.1.

[28] 张翔鹤, 王晓丽, 刘婷婷, 等. 桔小实蝇等六种常见果园害虫图像数据集[J]. 农业大数据学报, 2022, 4(1): 114-118.
ZHANG X H, WANG X L, LIU T T, et al. Image data set of six common orchard pests such as Bactrocera dorsalis[J]. Journal of agricultural big data, 2022, 4(1): 114-118.

[29] DING W G, TAYLOR G. Automatic moth detection from trap images for pest management[J]. Computers and electronics in agriculture, 2016, 123: 17-28.

[30] 徐小康. 图像目标数据集均衡完备构建技术研究[D]. 杭州: 杭州电子科技大学, 2021.
XU X K. Research on the balanced and complete construction technology of image target dataset[D]. Hangzhou: Hangzhou Dianzi University, 2021.

[31] ZHU M J, HAN K, WU E H, et al. Dynamic Resolution Network[EB/OL]. arXiv: 2106.02898, 2021.

[32] 周玉, 孙红玉, 房倩, 等. 不平衡数据集分类方法研究综述[J]. 计算机应用研究, 2022, 39(6): 1615-1621.
ZHOU Y, SUN H Y, FANG Q, et al. Review of imbalanced data classification methods[J]. Application research of computers, 2022, 39(6): 1615-1621.

[33] 林胜, 巩名轶, 牟文芊, 等. 基于对抗式生成网络的农作物病虫害图像扩充[J]. 电子技术与软件工程, 2020(3): 140-142.
LIN S, GONG M Y, MOU W Q, et al. Image expansion of crop diseases and pests based on antagonistic generation network[J]. Electronic technology & software engineering, 2020(3): 140-142.

[34] 史燕燕, 史殿习, 乔子腾, 等. 小样本目标检测研究综述[J]. 计算机学报, 2023, 46(8): 1753-1780.
SHI Y Y, SHI D X, QIAO Z T, et al. A survey on recent advances in few-shot object detection[J]. Chinese journal of computers, 2023, 46(8): 1753-1780.

[35] WANG J H, CHENG M M, JIANG J M. Domain shift preservation for zero-shot domain adaptation[J]. IEEE transactions on image processing: A publication of the IEEE Signal Processing Society, 2021, 30: 5505-5517.

[36] 汪启伟. 图像直方图特征及其应用研究[D]. 合肥: 中国科学技术大学, 2014.
WANG Q W. Study on image histogram feature and application[D]. Hefei: University of Science and Technology of China, 2014.

[37] HE Y, ZHOU Z Y, TIAN L H, et al. Brown rice planthopper (*Nilaparvata lugens* Stal) detection based on deep learning[J]. Precision agriculture, 2020, 21(6): 1385-1402.

[38] CHODEY M D, NOORULLAH SHARIFF C. Hybrid deep learning model for in-field pest detection on real-time field monitoring[J]. Journal of plant diseases and protection, 2022, 129(3): 635-650.

[39] 范馨月, 鲍泓, 潘卫国. 基于类别不平衡数据集的图像实例分割方法[J]. 计算机工程, 2022, 48(12): 224-231.
FAN X Y, BAO H, PAN W G. Image instance segmentation method based on class-imbalanced dataset[J]. Computer engineering, 2022, 48(12): 224-231.

[40] 刘浏. 基于深度学习的农作物害虫检测方法研究与应用[D]. 合肥: 中国科学技术大学, 2020.
LIU L. Research and applications on agricultural crop pest detection techniques based on deep learning[D]. Hefei: University of Science and Technology of China, 2020.

[41] JIAO L, DONG S F, ZHANG S Y, et al. AF-RCNN: An anchor-free convolutional neural network for multi-categories agricultural pest detection[J]. Computers and electronics in agriculture, 2020, 174: ID 105522.

[42] SHI Z C, DANG H, LIU Z C, et al. Detection and identification of stored-grain insects using deep learning: A more effective neural network[J]. IEEE access, 2020, 8: 163703-163714.

[43] 盛家文. 基于机器视觉的农业虫害测报研究[D]. 杭州: 浙江理工大学, 2020.
SHENG J W. Research on agricultural pest survey based on machine vision[D]. Hangzhou: Zhejiang Sci-Tech University, 2020.

[44] LI S Q, ZENG C, LIU S P, et al. Merging fixation for saliency detection in a multilayer graph[J]. Neurocomputing, 2017, 230: 173-183.

[45] LI W Y, WANG D J, LI M, et al. Field detection of tiny pests from sticky trap images using deep learning in agricultural greenhouse[J]. Computers and electronics in agriculture, 2021, 183: ID 106048.

[46] CHODEY M D, NOORULLAH SHARIFF C. Hybrid deep learning model for in-field pest detection on real-time field monitoring[J]. Journal of plant diseases and protection, 2022, 129(3): 635-650.

[47] DU J M, LIU L, LI R, et al. Towards densely clustered tiny pest detection in the wild environment[J]. Neurocomputing, 2022, 490: 400-412.

[48] BORJI A, SIHITE D N, ITTI L. Salient object detection: A benchmark[C]// European conference on computer vision. Berlin, German: Springer, 2012: 414-429.

[49] SUN C, SHRIVASTAVA A, SINGH S, et al. Revisiting unreasonable effectiveness of data in deep learning era[C]// 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, New Jersey, USA: IEEE, 2017: 843-852.

[50] 李楚为, 张志龙, 李树新. MTMS300: 面向显著物体检测的多目标多尺度基准数据集[J]. 中国图象图形学报, 2022, 27(4): 1039-1055.
LI C W, ZHANG Z L, LI S X. MTMS300: A multiple-targets and multiple-scales benchmark dataset for salient object detection[J]. Journal of image and graphics, 2022, 27 (4): 1039-1055.

[51] 王自全, 张永生, 于英, 等. 深度学习背景下视觉显著性物体检测综述[J]. 中国图象图形学报, 2022, 27(7): 2112-2128.
WANG Z Q, ZHANG Y S, YU Y, et al. Review of deep learning based salient object detection[J]. Journal of image and graphics, 2022, 27(7): 2112-2128.

[52] LI Y, HOU X D, KOCH C, et al. The secrets of salient object segmentation[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, New Jersey, USA: IEEE, 2014: 280-287.

[53] BYLINSKII Z, JUDD T, OLIVA A, et al. What do different evaluation metrics tell us about saliency models? [J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 41(3): 740-757.

[54] CHENG M M, MITRA N J, HUANG X L, et al. Global contrast based salient region detection[J]. IEEE transactions on pattern analysis and machine intelligence, 2015,

37(3): 569-582.

[55] BORJI A, CHENG M M, JIANG H Z, et al. Salient object detection: A benchmark[J]. IEEE transactions on image processing: A publication of the IEEE Signal Processing Society, 2015, 24(12): 5706-5722.

[56] FAN D P, CHENG M M, LIU J J, et al. Salient objects in clutter: Bringing salient object detection to the foreground[C]// European conference on computer vision. Berlin, German: Springer, 2018: 196-212.

[57] DEEPIKA P, KALIRAJ S. A survey on pest and disease monitoring of crops[C]// 2021 3rd International Conference on Signal Processing and Communication (ICPSC). Piscataway, New Jersey, USA: IEEE, 2021: 156-160.

[58] ZAYAS I Y, FLINN P W. Detection of insects in bulk wheat samples with machine vision[J]. Transactions of the ASAE, 1998, 41(3): 883-888.

[59] 韩瑞珍. 基于机器视觉的农田害虫快速检测与识别研究[D]. 杭州: 浙江大学, 2014.
HAN R Z. Research on fast detection and identification of field pests based on machine vision[D]. Hangzhou: Zhejiang University, 2014.

[60] 刘媛媛. 水稻害虫自动识别及分类系统[D]. 杭州: 中国计量大学, 2018.
LIU Y Y. Automatic identification and classification system for rice pests[D]. Hangzhou: China University of Metrology, 2018.

[61] WU H, WIESNER-HANKS T, STEWART E L, et al. Autonomous detection of plant disease symptoms directly from aerial imagery[J]. The plant phenome journal, 2019, 2(1): 1-9.

[62] 周瑶. 基于机器视觉与黄板诱导的有翅昆虫统计识别系统的研究与实现[D]. 重庆: 重庆大学, 2017.
ZHOU Y. Research and realization of winged insects' statistics and recognition system based on machine vision and yellow board induction[D]. Chongqing: Chongqing University, 2017.

[63] 白静亚. 基于机器视觉的棉田害虫图像采集与识别系统研究与改进[D]. 石河子: 石河子大学, 2022.
BAI J Y. Study and improvement on acquisition and recognition system of pest image in cotton field based on machine vision[D]. Shihezi: Shihezi University, 2022.

[64] LI D S, WANG R J, XIE C J, et al. A recognition method for rice plant diseases and pests video detection based on deep convolutional neural network[J]. Sensors, 2020, 20 (3): ID 578.

[65] PATIL R R, KUMAR S. Rice-fusion: A multimodality data fusion framework for rice disease diagnosis[J]. IEEE access, 2022, 10: 5207-5222.

[66] 史东旭, 高德民, 薛卫, 等. 基于物联网和大数据驱动的农业病虫害监测技术[J]. 南京农业大学学报, 2019, 42 (5): 967-974.
SHI D X, GAO D M, XUE W, et al. Research on agricultural disease and pest monitoring technology based on Internet of Things and big data[J]. Journal of Nanjing agricultural university, 2019, 42(5): 967-974.

[67] 张超, 王正, 姚青, 等. 便携式农业病虫害图像采集仪设计与应用[J]. 浙江农业科学, 2016, 57(12): 2077-2081.
ZHANG C, WANG Z, YAO Q, et al. Design and application of portable image acquisition instrument for agricultural pests and diseases[J]. Journal of Zhejiang agricultural sciences, 2016, 57(12): 2077-2081.

[68] HAWKESFORD M J, LORENCE A. Plant phenotyping: Increasing throughput and precision at multiple scales[J]. Functional plant biology, 2016, 44(1): v-vii.

[69] KAIVOSOJA J, HAUTSALO J, HEIKKINEN J, et al. Reference measurements in developing UAV systems for detecting pests, weeds, and diseases[J]. Remote sensing, 2021, 13(7): ID 1238.

[70] 蔡莉, 王淑婷, 刘俊晖, 等. 数据标注研究综述[J]. 软件学报, 2020, 31(2): 302-320.
CAI L, WANG S T, LIU J H, et al. Survey of data annotation[J]. Journal of software, 2020, 31(2): 302-320.

[71] ZHU J R, KAPLAN R, JOHNSON J, et al. HiDDeN: hiding data with deep networks[C]// European Conference on Computer Vision(ECCV). Berlin, German: Springer, 2018: 657-672.

[72] LEVINSON J, ASKELAND J, BECKER J, et al. Towards fully autonomous driving: Systems and algorithms[C]// 2011 IEEE Intelligent Vehicles Symposium (IV). Piscataway, New Jersey, USA: IEEE, 2011: 163-168.

[73] UIJLINGS J, KONYUSHKOVA K, LAMPERT C H, et al. Learning intelligent dialogs for bounding box annotation[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, New Jersey, USA: IEEE, 2018: 9175-9184.

[74] XIE C, MAO X Z, HUANG J J, et al. KOBAS 2.0: A web server for annotation and identification of enriched pathways and diseases[J]. Nucleic acids research, 2011, 39 (suppl_2): W316-W322.

[75] BERRIEL R F, ROSSI F S, DE SOUZA A F, et al. Automatic large-scale data acquisition via crowdsourcing for crosswalk classification: A deep learning approach[J]. Computers & graphics, 2017, 68: 32-42.

[76] HWANG M, JEONG Y, SUNG W K. Analysis of learning influence of training data selected by distribution consistency[J]. Sensors, 2021, 21(4): ID 1045.

[77] YUAN T T, DENG W H, TANG J A, et al. Signal-to-noise ratio: A robust distance metric for deep metric learning[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, New Jersey, USA: IEEE, 2019: 4815-4824.

[78] TAN B, ZHANG Y, PAN S J, et al. Distant domain transfer learning[C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. New York, USA: ACM, 2017: 2604-2610.

[79] STOJANOV P, GONG M M, CARBONELL J G, et al. Low-dimensional density ratio estimation for covariate shift correction[J]. Proceedings of machine learning research, 2019, 89: 3449-3458.

[80] BISHOP C M. Pattern recognition and machine learning[M]. New York: Springer, 2006.

[81] BLUMER A, EHRENFEUCHT A, HAUSSLER D, et al.

Learnability and the vapnik-chervonenkis dimension[J]. Journal of the ACM, 1989, 36(4): 929-965.
[82] BOSELLI R, CESARINI M, MERCORIO F, et al. An AI planning system for data cleaning[C]// Joint European conference on machine learning and knowledge discovery in databases. Cham, German: Springer, 2017: 349-353.
[83] GUPTA R, AUDHKHASI K, JACOKES Z, et al. Modeling multiple time series annotations as noisy distortions of the ground truth: An expectation-maximization approach[J]. IEEE transactions on affective computing, 2018, 9(1): 76-89.
[84] 曹伟. 众包域值标注算法研究[D]. 南京: 南京财经大学, 2016.
CAO W. Research of the algorithm of region-value annotation in crowdsouring[D]. Nanjing: Nanjing University of Finance and Economics, 2016.
[85] WANG Y W, RAO Y H, ZHAN X Y, et al. Sentiment and emotion classification over noisy labels[J]. Knowledge-based systems, 2016, 111: 207-216.
[86] RAYKAR V C, YU S P, ZHAO L H, et al. Supervised learning from multiple experts: Whom to trust when everyone lies a bit[C]// Proceedings of the 26th Annual International Conference on Machine Learning. New York, USA: ACM, 2009: 889-896.
[87] 于营, 杨婷婷, 杨博雄. 混淆矩阵分类性能评价及Python实现[J]. 现代计算机, 2021(20): 70-73, 79.
YU Y, YANG T T, YANG B X. Confusion matrix classification performance evaluation and python implementation[J]. Modern computer, 2021(20): 70-73, 79.
[88] RAYKAR VC, YU S, ZHAO L H, et al. Learning from crowds[J]. Journal of machine learning research, 2010, 11(2): 1297-1322.
[89] 于洪, 陈云. 基于Spark的三支聚类集成方法[J]. 郑州大学学报(理学版), 2018, 50(1): 20-26.
YU H, CHEN Y. Clustering ensemble method using three-way decisions based on spark[J]. Journal of Zhengzhou university (natural science edition), 2018, 50(1): 20-26.
[90] VOGEL T, HEISE A, DRAISBACH U, et al. Reach for gold: An annealing standard to evaluate duplicate detection results[J]. Journal of data and information quality, 2014, 5(1):1-25.
[91] KOLESNIKOV A, BEYER L, ZHAI X H, et al. Big transfer (BiT): General visual representation learning[M]// Computer vision ‑ ECCV 2020. Cham: Springer International Publishing, 2020: 491-507.
[92] XIA C Q, LI J, CHEN X W, et al. What is and what is not a salient object? Learning salient object detector by ensembling linear exemplar regressors[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, New Jersey, USA: IEEE, 2017: 4399-4407.
[93] 蒋心璐, 陈天恩, 王聪, 等. 农业害虫检测的深度学习算法综述[J]. 计算机工程与应用, 2023, 59(6): 30-44.
JIANG X L, CHEN Tian'en, WANG C, et al. Survey of deep learning algorithms for agricultural pest detection[J]. Computer engineering and applications, 2023, 59(6): 30-44.

The Key Issues and Evaluation Methods for Constructing Agricultural Pest and Disease Image Datasets: A Review

GUAN Bolun, ZHANG Liping, ZHU Jingbo, LI Runmei, KONG Juanjuan, WANG Yan, DONG Wei*

(*Institute of Agricultural Economy and Information*, *Anhui Academy of Agricultural Sciences*, *Hefei* 230001, *China*)

Abstract:

[Significance] The scientific dataset of agricultural pests and diseases is the foundation for monitoring and warning of agricultural pests and diseases. It is of great significance for the development of agricultural pest control, and is an important component of developing smart agriculture. The quality of the dataset affecting the effectiveness of image recognition algorithms, with the discovery of the importance of deep learning technology in intelligent monitoring of agricultural pests and diseases. The construction of high-quality agricultural pest and disease datasets is gradually attracting attention from scholars in this field. In the task of image recognition, on one hand, the recognition effect depends on the improvement strategy of the algorithm, and on the other hand, it depends on the quality of the dataset. The same recognition algorithm learns different features in different quality datasets, so its recognition performance also varies. In order to propose a dataset evaluation index to measure the quality of agricultural pest and disease datasets, this article analyzes the existing datasets and takes the challenges faced in constructing agricultural pest and disease image datasets as the starting point to review the construction of agricultural pest and disease datasets.

[Progress] Firstly, disease and pest datasets are divided into two categories: private datasets and public datasets. Private datasets have the characteristics of high annotation quality, high image quality, and a large number of inter class samples that are not publicly available. Public datasets have the characteristics of multiple types, low image quality, and poor annotation quality. Secondly, the problems faced in the construction process of datasets are summarized, including imbalanced categories at the dataset level, difficulty in feature extraction at the dataset sample level, and difficulty in measuring the dataset size at the usage level. These include imbalanced inter class and intra class samples, selection bias, multi-scale targets, dense targets, uneven data distribution, uneven image quality, insufficient dataset size, and dataset availability. The main reasons for the problem are analyzed by two key aspects of image acquisition and annotation methods in dataset construction, and the improvement strategies and suggestions for the algorithm to address the above issues are summarized. The collection devices of the dataset can be divided into handheld devices, drone platforms, and fixed collection devices. The collection method of handheld devices is flexible and convenient, but it is inefficient and requires high photography skills. The drone platform acquisition method is suitable for data collection in contiguous areas, but the detailed features captured are not clear enough. The fixed device acquisition method has higher efficiency, but the shooting scene is often relatively fixed. The annotation of image data is divided into rectangular annotation and polygonal annotation. In image recognition and detection, rectangular annotation is generally used more frequently. It is difficult to label images that are difficult to separate the target and background. Improper annotation can lead to the introduction of more noise or incomplete algorithm feature extraction. In response to the problems in the above three aspects, the evaluation methods are summarized for data distribution consistency, dataset size, and image annotation quality at the end of the article.

[Conclusions and Prospects] The future research and development suggestions for constructing high-quality agricultural pest and disease image datasets based are proposed on the actual needs of agricultural pest and disease image recognition:(1) Construct agricultural pest and disease datasets combined with practical usage scenarios. In order to enable the algorithm to extract richer target features, image data can be collected from multiple perspectives and environments to construct a dataset. According to actual needs, data categories can be scientifically and reasonably divided from the perspective of algorithm feature extraction, avoiding unreasonable inter class and intra class distances, and thus constructing a dataset that meets task requirements for classification and balanced feature distribution. (2) Balancing the relationship between datasets and algorithms. When improving algorithms, consider the more sufficient distribution of categories and features in the dataset, as well as the size of the dataset that matches the model, to improve algorithm accuracy, robustness, and practicality. It ensures that comparative experiments are conducted on algorithm improvement under the same evaluation standard dataset, and improved the pest and disease image recognition algorithm. Research the correlation between the scale of agricultural pest and disease image data and algorithm performance, study the relationship between data annotation methods and algorithms that are difficult to annotate pest and disease images, integrate recognition algorithms for fuzzy, dense, occluded targets, and propose evaluation indicators for agricultural pest and disease datasets. (3) Enhancing the use value of datasets. Datasets can not only be used for research on image recognition, but also for research on other business needs. The identification, collection, and annotation of target images is a challenging task in the construction process of pest and disease datasets. In the process of collecting image data, in addition to collecting images, attention can be paid to the collection of surrounding environmental information and host information. This method is used to construct a multimodal agricultural pest and disease dataset, fully leveraging the value of the dataset. In order to focus researchers on business innovation research, it is necessary to innovate the organizational form of data collection, develop a big data platform for agricultural diseases and pests, explore the correlation between multimodal data, improve the accessibility and convenience of data, and provide efficient services for application implementation and business innovation.

Key words: agricultural pests; data set; deep learning; monitoring and warning; data acquisition; data annotations; data set evaluation

Foundation items: General Program of National Natural Science Foundation of China (32171888); Anhui Province Financial Agricultural Scientific and Technological Achievements Transformation Project (2022ZH001); Anhui Academy of Agricultural Sciences Research Platform Project(2023YL014)